DOCUMENT RESUME

ED 173 214 SO 011 772

AUTHOR Tuma, Nancy Brandon
TITLE Final Report for Dynamic Models for Causal Analysis of Panel Data. When Can Interdependence in A Dynamic System of Qualitative Variables Be Ignored? Part III, Chapter 8.
INSTITUTION Stanford Univ., Calif.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.; National Inst. of Mental Health (DHEW), Bethesda, Md.
PUB DATE Dec 78
GRANT NIE-G-76-0082; NIMH-5-R01-MH29208-02
NOTE 48p.; Tables 1, 3-4 may not reproduce clearly in paper copy from EDRS due to small print type of original document; For related documents, see SO 011 759-772; An earlier version of this paper was presented at the Mathematical Social Board Advanced Research Symposium on Stochastic Process Models of Social Structure (Pittsburgh, Pennsylvania, December 1-2, 1977)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Data Analysis; *Educational Change; Measurement Techniques; *Research Design; *Research Methodology; *Research Problems; *Social Change; Social Science Research; Sociometric Techniques; Statistical Bias; *Validity

ABSTRACT

This document is part of a series of chapters described in SO 011 759. This chapter offers guidelines for studying change in dynamically interdependent variables, even when a model that ignores interdependence is estimated. Many sociological studies concern dynamically interdependent variables; for example, in a study of female employment, a woman's employment may influence and be influenced by her marital status. The report addresses these issues: Under what circumstances do inferences based on a misspecified model (one that ignores interdependence among variables) lead to errors that are comparatively small and what kinds of research designs minimize errors and what kinds accentuate them? Separate sections examine assumptions of models, implications of bivariate and univariate models, research design, and procedures and findings for the 23 bivariate processes considered. Results indicate that biases from ignoring dynamic interrelationships tend to increase with the degree of interdependence of the variables. However, under some circumstances, biases are not large when interdependencies are ignored. If the variable being studied changes more slowly than the ignored variable, biases are relatively small. For example, research on marital stability may ignore changes in employment status, but not the reverse. Advantages of using event-history and panel analyses are also discussed. (Author/KC)

ED173214

APR 5 1979

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Final Report for

DYNAMIC MODELS FOR CAUSAL ANALYSIS OF PANEL DATA

WHEN CAN INTERDEPENDENCE IN A DYNAMIC SYSTEM OF QUALITATIVE VARIABLES BE IGNORED?*

by

Nancy Brandon Tuma
Stanford University

November, 1977
Revised March, 1978
Revised and Retitled December, 1978

SE 011 772

*I wish to thank Marlos Viana for assistance in deriving some of the equations in this paper and Glenn Carroll for assistance with programming. I appreciate helpful comments on an earlier version given by Michael T. Hannan, John Lehoczky, Karl Schuessler, Joseph Schwartz, Burton Singer, Aage B. Sørensen, Barbara Warsavage, and Stanley S. Wasserman. This research was supported by National Institute of Education Grant #NIE-G-76-0082 and by National Institute of Mental Health Grant #NIMH-5-RO1-MH29208-02. An earlier version of this paper was presented at the Mathematical Social Board Advanced Research Symposium on Stochastic Process Models of Social Structure, in Pittsburgh, Pennsylvania, December 1-2, 1977.

Many phenomena of interest to sociologists concern variables that are dynamically interdependent with other variables. For example, we may wish to understand growth in economic productivity of nations, and a nation's economic growth may depend on its investments in education, while at the same time growth in educational investments may depend on economic productivity. Or, we may want to explain female employment, and a woman's employment may influence, and be influenced by, her marital status.

Although variables are dynamically interdependent, we often study change in one (or more) of these variables as though certain interrelationships do not exist. Sometimes we lack information on some variables thought to be interrelated with other observed variables. Faced with incomplete data, we may choose to ignore the interdependencies between observed and unobserved variables in the models actually estimated.[1] Sometimes we have the necessary data but believe certain interrelationships are so weak that ignoring them will have little practical impact on our conclusions. And sometimes--undoubtedly all too often--we simply do not realize that certain variables have an important dynamic interdependency with others.

Whatever the reason, we often estimate the wrong model--in particular, a model that ignores the dynamic interdependence among variables. Given this fact, we would like to know the following: (1) Under what circumstances do inferences based on this kind of misspecified model lead to errors that are comparatively small? (2) What kinds of research designs minimize errors, and what kinds accentuate them?

There is a well-developed body of literature on these issues when

the variables are quantitative, as in the first of the two examples given above.[2] Furthermore, in the past decade sociologists have become increasingly sensitive to these problems and skillful in developing data and analytic designs that attempt to mitigate them.

I focus on these issues when the variables are qualitative, as in the second example above. First, there is not yet a large stock of classic works to which sociologists can turn when investigating qualitative variables with dynamic interdependencies. Second, sensitivity to the consequences of model misspecification still seems to be low in studies of qualitative variables. For example, marital status and employment status are rarely treated as a system of endogenous variables. Yet, it is extremely common to control marital status in studies of employment behavior of men and women, and also common to control employment status (or variables related to it) in studies of marital stability. It implicitly seems to be the view that controlling for the value of a qualitative variable at one time deals adequately with the dynamic interdependence among qualitative variables. Such folk practices may be good ones, but evidence supporting their value is lacking.

The main objectives of this paper are to heighten awareness of the problem of ignoring interdependence in a dynamic system of qualitative variables and to stimulate inquiry into consequences of this type of model misspecification. An additional objective is to suggest tentative answers to the two questions raised above. These tentative conclusions are based on a study of illustrative cases. More rigorous investigation is necessary to establish their generality.

In the particular cases examined in this chapter I assume that a time-homogeneous, continuous-time, discrete-state Markov process describes the interrelationship among a collection of dichotomous variables. A Markov process allows change in one variable to depend on its own level and also on the level of other variables, though not on previous history. From a mathematical viewpoint Markov processes are comparatively simple, which facilitates the study of model misspecification. Even for this restricted class of models, it is difficult to study the consequences of model misspecification when the number of interdependent variables is large. Consequently, I focus on a bivariate Markov process $Y(t) = \{Y_1(t), Y_2(t)\}$, where $Y_1(t)$ and $Y_2(t)$ are dichotomous random variables that vary over time t, as well as over members of a population. I suppose that the univariate Markov process $Y_1(t)$ forms the basis of the analysis, rather than the true bivariate process.

I summarize the model assumptions in Section I and report various implications of the bivariate and univariate Markov processes in Section II. In Section III I describe and give estimation equations for several research designs that might be used to estimate parameters in the univariate model of $Y_1(t)$. I consider designs in which there are two-wave panel data on $Y_1(0)$ and $Y_1(t)$ and event-history data on $Y_1(t)$ in the given interval $0 \leq t \leq T$. For each kind of data on Y_1, I consider two types of sample designs. In one design the investigator has a random sample from a population in equilibrium so that $Y_2(0)$ has an equilibrium probability distribution. In the other design the investigator selects a sample to be homogeneous on $Y_2(0)$, presumably to control for a covariate thought to be related to Y_1.

Section IV describes the procedure used to study the consequences of ignoring dynamic interdependencies for different outcomes when different research designs are utilized. Because the equations describing the consequences of this kind of model misspecification are very complicated even for a bivariate Markov process, I consider asymptotic results for specific cases selected to resemble phenomena of substantive interest. Section IV contains a summary of the findings for the twenty-eight bivariate processes that I considered.

The most provocative finding is that biases are very small for all research designs and outcomes when Y_1 changes much more slowly than Y_2. On the other hand, biases are often large when Y_1 does change much more rapidly than Y_2, although event-history analysis of a random sample tends to lead to smaller biases than other research designs. An illustrative substantive implication of this finding is that an analysis of marital stability which ignores the interdependence between marital status and employment status should still lead to accurate descriptions of marital behavior. This is because most people change marital partners much less often than their employment status. On the other hand, studies of employment behavior that ignore the interdependence between marital status and employment status are likely to give poor predictions about employment behavior.

I. ASSUMPTIONS

I confine my attention to continuous-time, discrete-state Markov models of change in interdependent dichotomous variables. Suppose that the system of interest consists of M binary-valued random variables that vary over time t, $Y_m(t)$, m=1,..,M, where the state space of $Y_m(t)$ is $S_m=\{0,1\}$. Description of change in the system is made more compact by defining a new random variable Y(t) as the following linear function of these M random variables:[3]

$$Y(t) \triangleq \sum_{m=1}^{M} 2^{(M-m)} Y_m(t) \qquad (1)$$

The state space S of Y(t) consists of the integers from 0 to 2^M-1 and so has 2^M elements. In creating Y(t), I have essentially just treated the ordered M-tuples of the values of the set of random variables, $(y_1(t),..y_M(t))$, as binary numbers and converted them to a base ten number system.

It is mathematically convenient to assume that the process describing Y(t) is Markovian. There are many excellent statements of the assumptions of the continuous-time, discrete-state Markov process (e.g., Cox and Miller, 1965; Feller, 1968). I summarize these to establish my notation.

Let s and y be any states in S, and let Y(u) = s and Y(t) = y. In addition, let $p_{sy}(u,t)$ represent the probability of a transition from state s at time u to state y at time t, $u \leq t$:

$$p_{sy}(u,t) \triangleq \text{Prob}[Y(t) = y \mid Y(u) = s] \qquad (2)$$

Let P(u,t) denote the matrix of the transition probabilities: $P(u,t) = \{p_{sy}(u,t)\}$.

Let $r_{ys}(t)$ be the instantaneous rate of a transition at time t from state y to any other state s, which is defined as the limit, as Δt approaches zero, of the probability of a transition from y to s in the interval between t and $t+\Delta t$:

$$r_{ys}(t) \triangleq \lim_{\Delta t \to 0} \frac{p_{ys}(t,t+\Delta t)}{\Delta t}, \qquad y \neq s \qquad (3)$$

Note that a transition rate cannot be negative because it is a probability measure (per unit of time).[4] I also assume that all transition rates in

the system are finite; this is a reasonable assumption in social systems. The definition of the rate of leaving state y at time t, $r_y(t) \triangleq -r_{yy}(t)$, is:

$$r_y(t) \triangleq -r_{yy}(t) = \sum_{\substack{s \\ s \neq y}} r_{ys}(t) \quad . \tag{4}$$

Because of (4), a Markov process with 2^M states has at most $2^M(2^M-1)$ unique transition rates. For example, for a model of change in the single dichotomous variable $Y_1(t)$, the two rates $r_{01}(t)$ and $r_{10}(t)$ are sufficient to define change in $Y_1(t)$ over time.

The Markov assumption implies that the Chapman-Kolmogorov identity holds:

$$P(v,t) = P(v,u)\, P(u,t), \qquad v \leq u \leq t \; . \tag{5}$$

Given this assumption and other natural ones (e.g., continuity, probabilities ranging from zero to one, etc.),[5] it can then be shown that

$$\frac{dP(u,t)}{dt} = P(u,t)\, R(t) \tag{6}$$

where R(t) is a matrix in which the y^{th} diagonal element is $r_y(t)$ and the ys^{th} off-diagonal element is $r_{ys}(t)$.

In practice it is usually assumed that transition rates are homogeneous within some specified population and do not vary over time. Neither population homogeneity nor time-stationarity of transition rates is a very realistic assumption insofar as change in a social system is concerned.[6] Nevertheless, I maintain these assumptions in this paper to simplify the discussion of problems arising from model misspecification.

I also make a further simplifying assumption--namely, that only one of the M variables may change in any sufficiently small moment of time. This simplification, which many other investigators have made (see, for example, Coleman, 1964; Sørensen and Hallinan, 1975; Holland and Leinhardt, 1977), is frequently a reasonable approximation to the actual process of change in a social system. Its primary methodological advantage is that it reduces the number of unique, nonzero transition rates that must be estimated from $2^M(2^M-1)$ to $2^M M$. This is a drastic decrease even when M is moderate in size. For two variables, there are eight rather than twelve rates to be estimated; for three variables, there are twenty-four rather than fifty-six rates; for four variables, sixty-four rather than 240. Yet $2^M M$ is large even when M is fairly small. Consequently, we often introduce still more simplifying assumptions.

One such simplification is to ignore some variables. Most variables of sociological interest are interrelated to other variables. Ordinarily we cannot study change in all variables that we believe to be interdependent (in a nontrivial way), either because of data limitations or because of insufficient resources for conducting such an analysis. So it is important to understand what happens when we study change in some variables and ignore change in others. If possible, we would like to know how various features of the true process and of our research design affect the size and direction of errors that arise from this procedure. Then we may be able to judge when we can obtain sensible conclusions from a misspecified model, and perhaps even learn how to adjust our conclusions for this.

It is difficult (perhaps impossible) to obtain a general statement of the consequences of this type of model misspecification.

For this reason, and for expository purposes, I concentrate on the simplest case--that of two interdependent variables. This case can be applied to a number of sociologically interesting problems. Furthermore, understanding it may help us make intelligent guesses about the consequences of ignoring dynamic interdependencies in systems containing a large number of variables.

In formal terms, I assume that the true situation is described by a time-stationary Markov model of change in <u>two</u> dichotomous variables, $Y_1(t)$ and $Y_2(t)$, that cannot change simultaneously. This model is depicted in Figure 1a. Its matrix of transition rates is:

$$R = \begin{bmatrix} -r_0 & r_{01} & r_{02} & 0 \\ r_{10} & -r_1 & 0 & r_{13} \\ r_{20} & 0 & -r_2 & r_{23} \\ 0 & r_{31} & r_{32} & -r_3 \end{bmatrix} \tag{7}$$

where $r_y \overset{\Delta}{=} \sum_{\substack{s \\ s \neq y}} r_{ys}$, for y = 0, 1, 2, 3.

I suppose that a time-stationary Markov model of change in <u>only one</u> of the two variables is actually estimated. For convenience I assume that $Y_1(t)$ is studied and $Y_2(t)$ is ignored. The univariate model of change in $Y_1(t)$ is shown in Figure 1b. It has the matrix of transition rates

$$R^* = \begin{bmatrix} -r_0^* & r_{01}^* \\ r_{10}^* & -r_1^* \end{bmatrix} \tag{8}$$

where $r_y^* \triangleq r_{ys}^*$, for y = 0, 1 and s ≠ y. Throughout the chapter a "*" as a superscript on a symbol refers to the univariate model.

When $r_{01} = r_{23}$, $r_{02} = r_{13}$, $r_{10} = r_{32}$, and $r_{20} = r_{31}$, the rate of change in the value of Y_1 is independent of the value of Y_2, and the rate of change in the value of Y_2 is independent of the value of Y_1. If this is true, implications of (8) for outcomes related to Y_1 will be identical to the implications of (7). But when these identities do not hold and Y_1 and Y_2 are dynamically interdependent, (7) and (8) have different implications, as outlined below.

II. SOME IMPLICATIONS OF THE BIVARIATE AND UNIVARIATE MODELS

Sociologists may wish to summarize what they know about change in a qualitative variable in terms of three variables: the probability that the variable has a particular value at some point in time; the length of time that the variable (continuously) has one value rather than another; and the number of times that the variable changes its value in some period of time. A Markov model has implications for each of these outcomes. The implications of the bivariate model for these quantities let us predict the true expected value of each. The implications of the univariate model tell us what we will actually predict assuming we have estimated r_{01}^* and r_{10}^* rather than the rates in (7). Below I give the solutions for each outcome for both the bivariate and univariate models, but I do not attempt to interpret these complex equations directly.

Probability Distribution

Usually one question that we would like to answer is: What is the

probability that the variable being studied has a given value at some given point in time? For example, we might want to know the probability that a person is married, working, or friends with someone else at a particular moment. Or, we might want to know the probability that two nations have diplomatic ties.

Formally this asks what is $p_{y_1}(t)$, the probability that $Y_1(t) = y_1$. Given the assumptions of a Markov model, we can readily derive this probability. In general it depends on time t and the initial distribution. To avoid arbitrary assumptions about the latter two quantities, I focus on the equilibrum (steady-state) probability, $p_{y_1}(\infty)$, the value of $p_{y_1}(t)$ as t approaches infinity.[7]

For the univariate model in (8),

$$p_1^*(\infty) = 1 - p_0^*(\infty) = r_{01}^*/(r_{01}^* + r_{10}^*) \tag{9}$$

where $p_{y_1}^*(\infty)$ is the equilibrium probability that $Y_1(\infty) = y_1$. For the bivariate model in (7), it can be shown that for $y_1 = 1$:

$$\begin{aligned} p_{y_1}(\infty) = \{&r_{01}r_{13}(r_2 + r_{32}) + r_{02}[r_{23}(r_1 + r_{32}) \\ &+ r_{01}r_{10} + r_{02}r_{20} + (r_0 + \lambda_2)(r_0 + \lambda_3) \\ &+ (r_0 + r_2 + \lambda_2 + \lambda_3)(r_2 + \lambda_4)]\}/(-\lambda_2\lambda_3\lambda_4) \end{aligned} \tag{10}$$

where λ_j, $j = 2,3,4$, are distinct, negative, nonzero eigenvalues of R in (8).[8]

Duration

The duration between successive changes in the value of Y_1 adds a temporal dimension to our knowledge about a process. For example, we might wish to know the length of time that a person is employed or unemployed, married or single, or friends with another person. Or, we might want to know the length of time that two countries continue diplomatic relations with one another, or that a union and a firm have gone without a contract.

Even though every member of a population has the same transition rates, the duration in a state has a distribution with a nonzero variance, given the assumptions of a Markov model. I limit consideration to D_{y_1}, the mean (or expected) duration that $Y_1(t)$ continuously equals y_1. For the univariate model,

$$D^*_{y_1} = 1/r^*_{y_1} \tag{11}$$

For the bivariate model,

$$D_{y_1} = \frac{\rho_{y_1}(r_b + r_{ab}) + (1-\rho_{y_1})(r_a + r_{ba})}{r_a r_b - r_{ab} r_{ba}} \tag{12}$$

where $a = 2y_1 + 1$, $b = 2y_1$, and ρ_{y_1} is the probability that $Y_2(0) = 1$ when $Y_1(0) = y_1$.

Number of Changes in a Period

A third way of summarizing information about the change process is in terms of the number of times a particular kind of change occurs in a given interval for a particular member of the population being studied. For

example, an investigator studying marital stability might be concerned with the number of marriages or the number of marital dissolutions that a person has within some period of time. Someone studying work behavior might want to know the number of different jobs that a person holds in a given interval of his or her career. Others might be interested in the number of times two countries break off diplomatic relations during a given historical period.

As in the case of duration, a Markov model of change in a set of dichotomous variables implies that the number of times any given variable changes in a period has a distribution with a nonzero variance, even though everyone in the population has the same transition rates. However, except for very simple Markov models (e.g., a Poisson model), it is difficult (and sometimes impossible) to write this distribution as an explicit function of the transition rates. I concentrate on the expected number, a useful summary statistic that can be derived more readily than the probability distribution.

For the univariate model, the expected number of changes to s before time t when state y_1 is occupied at time 0, $n^*_{y_1 s}(t)$, is:

$$n^*_{y_1 y_1}(t) = \frac{r^*_0 r^*_1 t}{(r^*_0+r^*_1)} + \frac{r^*_0 r^*_1}{(r^*_0+r^*_1)^2}\left[e^{-(r^*_0+r^*_1)t} - 1\right] \tag{13}$$

$$n^*_{y_1(1-y_1)}(t) = \frac{r^*_0 r^*_1 t}{(r^*_0+r^*_1)} - \frac{(r^*_{(1-y)y})^2}{(r^*_0+r^*_1)^2}\left[e^{-(r^*_0+r^*_1)t} - 1\right] \tag{14}$$

The derivation of the equations for the bivariate model requires an additional assumption that has not previously been necessary--namely, that the process is in equilibrium. Because of its length, the solution is given in the appendix.

III. RESEARCH DESIGN

We would like to choose a research design so that inferences from a sample will be as close as possible to the true values in the population, i.e., errors will be small on the average. Choice of a design involves many decisions, most of which I ignore. In particular, I disregard effects of the sample size N, assuming that N is so large that variations in it do not affect conclusions. Instead, I focus on two decisions that seem especially problematic when estimating a misspecified model. First, what kind of data on the variable(s) being explained should be used to estimate parameters? Second, what kind of sample design should be used? In particular, should we stratify on values of the omitted endogenous variable(s), such as Y_2, or should we simply sample at random, ignoring the omitted endogenous variable(s) entirely?

I consider estimation from two kinds of data on Y_1: two-wave panel data, in which we observe $Y_1(0)$ and $Y_1(t)$ for every member of a sample for given times 0 and t, and event-history data, in which we observe $Y_1(t)$ for all t in a given time interval $0 \leq t \leq T$.[9] For a correctly specified model, event-history analysis is superior to panel analysis (Tuma et al., 1979), but this advantage may not remain when the model is misspecified.

I consider two kinds of samples: a sample selected so that everyone has the same value on Y_2 initially (a "stratified" sample), and a sample

selected randomly from a population in equilibrium (a "random" or "type r" sample). Stratified samples include those in which $Y_2(0) = 0$ for all N sample members (a "type 0" sample), and those in which $Y_2(0) = 1$ for all N sample members (a "type 1" sample). Obviously ρ_{y_1}, the probability that $Y_2(0) = 1$ when $Y_1(0) = y_1$ (which affects estimates of rates in the misspecified, univariate model), is zero in a type 0 sample and unity in a type 1 sample. In the random or type r sample,

$$\rho_0 = p_1(\infty)/[p_0(\infty) + p_1(\infty)] \tag{15}$$

$$\rho_1 = [p_3(\infty)/ p_2(\infty) + p_3(\infty)] \tag{16}$$

and

$$p_0(\infty) = [(r_0 + \lambda_2)(r_0 + \lambda_3)(r_0 + \lambda_4) + r_{01}r_{10}(r_0 - r_2 - r_3) + r_{02}r_{20}(r_0 - r_1 - r_4)]/(\lambda_2\lambda_3\lambda_4) \tag{17}$$

$$p_1(\infty) = \{r_{01}[(r_0 + \lambda_2)(r_0 + \lambda_3) + (r_0 + r_1 + \lambda_2 + \lambda_3)(r_1 + \lambda_4) + r_{01}r_{10} + r_{02}r_{20} + r_{13}r_{31}] + r_{02}r_{23}r_{31}\}/(-\lambda_2\lambda_3\lambda_4) \tag{18}$$

$$p_2(\infty) = \{r_{02}[(r_0 + \lambda_2)(r_0 + \lambda_3) + (r_0 + r_2 + \lambda_2 + \lambda_3)(r_2 + \lambda_4) + r_{01}r_{10} + r_{02}r_{20} + r_{23}r_{32}] + r_{01}r_{13}r_{32}\}/(-\lambda_2\lambda_3\lambda_4) \tag{19}$$

$$p_3(\infty) = [r_{01}r_{13}r_2 + r_{02}r_{23}r_1]/(-\lambda_2\lambda_3\lambda_4) \tag{20}$$

where $p_y(\infty)$ is the steady-state probability that $Y(t) = y$, and λ_j, $j = 2, 3, 4$, continue to represent the district, negative, nonzero eigenvalues of R in (7). (Note that the sum of (19) and (20) gives (11), reported earlier.)

It may seem better to analyze type 0 and type 1 samples separately than to analyze a random or type r sample. Unlike the latter procedure, the former partially controls for the omitted endogenous variable Y_2. Ordinarily we expect such controls to give improved estimates. However, the findings below indicate that the expected improvement does not always occur.

Two-Wave Panel Data

Data from a two-wave panel provide information on $Y_1(0)$ and $Y_1(t)$ for the N members of the sample. The transition probabilities in the univariate model can be estimated from these data as:

$$\hat{p}^*_{ik}(t) = \frac{\#\text{ for whom } Y_1(0) = i \text{ and } Y_1(t) = k}{\#\text{ for whom } Y_1(0) = i} \tag{21}$$

for i, k = 0,1. When

$$\hat{p}^*_{00}(0,t) + \hat{p}^*_{11}(0,t) > 1 \tag{22}$$

the unique maximum likelihood estimators of the rates r^*_{01} and r^*_{10} are as follows (Singer and Spilerman, 1974, p. 392):[10]

$$\hat{r}^*_{01} = \frac{\hat{p}^*_{01}(0,t) \ln [\hat{p}^*_{00}(0,t) + \hat{p}^*_{11}(0,t) - 1]}{t [\hat{p}^*_{00}(0,t) + \hat{p}^*_{11}(0,t) - 2]} \tag{23}$$

$$\hat{r}^*_{10} = \frac{\hat{p}^*_{10}(0,t) \ln [\hat{p}^*_{00}(0,t) + \hat{p}^*_{11}(0,t) - 1]}{t [\hat{p}^*_{00}(0,t) + \hat{p}^*_{11}(0,t) - 2]} \tag{24}$$

To relate the estimates of the rates in the univariate model to the (true) rates in the bivariate process, we must express $p^{*}_{ik}(0,t)$ as a function of the rates for the bivariate process. When the sample size N is large, as assumed here, the sample estimates of $p^{*}_{ik}(0,t)$ approximately equals its true value. The latter is a function of ρ_i, the probability that $Y_2(0) = 1$ given that $Y_1(0) = i$ (which depends on the type of sample), and the matrix of the true probabilities of a change from $Y(0) = s$ to $Y(t) = y$, $p_{sy}(t)$:

$$p^{*}_{01}(0,t) = (1-\rho_0)[p_{02}(0,t) + p_{03}(0,t)] + \rho_0[p_{12}(0,t) + p_{13}(0,t)] \tag{25}$$

$$p^{*}_{10}(0,t) = (1-\rho_1)[p_{20}(0,t) + p_{21}(0,t)] + \rho_1[p_{30}(0,t) + p_{31}(0,t)] \tag{26}$$

where $p_{sy}(0,t)$ is an element of the matrix $P(0,t)$ given by:[11]

$$P(0,t) = \sum_{j=1}^{2^M} A_j e^{t\lambda_j} \tag{27}$$

$$A_j = \prod_{\substack{k \\ j \neq k}} (R - \lambda_k)/(\lambda_j - \lambda_k) \tag{28}$$

Event-History Data

We can also estimate transition rates from event-history data, which record $Y_1(t)$ for $0 \leq t \leq T$ for all N sample members. From such data we can obtain information on each sample member on the values of the random

variables $n_{y_1}(T)$, the number of changes to y_1 in the interval (0,T), and $U_{y_1}(T)$, the total time that $Y_1(t) = y_1$ in (0,T). The maximum likelihood estimates of r^*_{01} and r^*_{10} are then calculated as follows (Albert, 1962):

$$\hat{r}^*_{01} = [\sum_{i=1}^{N} n_1(T)]/[\sum_{i=1}^{N} U_0(T)] \tag{29}$$

$$\hat{r}^*_{10} = [\sum_{i=1}^{N} n_0(T)]/[\sum_{i=1}^{N} U_1(T)] \tag{30}$$

The numerators and denominators of the right-hand sides of equations (29) and (30) are random variables and thus vary from one sample to another. When the sample size N and the length of the observation period T are sufficiently large, as assumed here, the right-hand side of each of these equations equals its expected value in the population, i.e.,

$$\hat{r}^*_{01} = 1/D_0 \tag{31}$$

$$\hat{r}^*_{10} = 1/D_1 \tag{32}$$

To relate $\hat{r}^*_{01}$ and $\hat{r}^*_{10}$ to the rates of the bivariate process we need only substitute into (31) and (32) the values of D_0 and D_1, which are given by (12). Note that D_0 and D_1 depend on ρ_0 and ρ_1, as well as the transition rates of the bivariate process.

IV. PROCEDURE

In a correctly specified model, the expected error (or bias) in parameter

estimates and in outcomes predicted from such estimates--the average difference between the inference from a sample and the true population value--is zero for the outcomes and research designs discussed in the previous two sections.[12] In a misspecified model, the bias is generally nonzero. When the estimated model consists of a univariate Markov process and a bivariate Markov process actually describes change in the system, bias is a complicated nonlinear function of the eight, unique, nonzero transition rates of the bivariate process given in (7). The bias also depends on the particular outcome considered and on the research design--the kind of data collected and the type of sample. We can use the equations in the previous two sections to write equations for the bias, but these equations are so complex that they are not readily interpretable.

Because of this complexity, I decided not to study bias analytically, but instead to calculate it for selected cases. The latter approach has an important disadvantage: we cannot be certain of obtaining similar conclusions about patterns of bias if the true process differs markedly from those studied. However, if we refrain from generalizing to very different kinds of processes, it also has an important advantage: we can see the size and direction of biases for cases similar to those of particular interest.

I chose to study specific cases that resemble two phenomena of substantive interest to me: the dynamic interrelationship between employment status and marital status of men and women, and the interdependence of changes in children's friendships. Partly because of data limitations and partly to avoid complicated analytical problems associated with modeling dynamic interdependencies between qualitative variables, my

previous empirical research in these areas was based on models that ignored dynamic interdependencies. I undertook the research discussed in this chapter to assess the extent to which my previous research in these substantive areas may have erred because of this type of model misspecification.

I examined four bivariate systems or pairs of variables. Each pair consists of two dichotomous variables, Y_A and Y_B. For each pair I consider the bias arising from studying Y_A (i.e., setting $Y_1 = Y_A$) and ignoring Y_B(i.e., setting $Y_2 = Y_B$) and from studying Y_B and ignoring Y_A. For reasons that become apparent later, Y_A always denotes the more slowly changing variable in the pair; that is, Y_A and Y_B are always defined so that the number of changes in Y_A in any interval (0,T) is less than the number of changes in Y_B in the same interval (0,T).

The four pairs are as follows. Pair 1 refers to a man's marital status (Y_A) and his employment status (Y_B); Pair 2 refers to a woman's marital status (Y_A) and her employment status (Y_B). Pair 3 refers to a wife's employment status (Y_A) and her husband's employment status (Y_B). Pair 4 refers to person A's choice of person B as a friend (Y_A) and person B's choice of person A as a friend (Y_B).

Before calculating biases from estimating a misspecified univariate model, we must first select values for the transition rates in the true bivariate process. I began by assuming that variables were independent and then choosing transition rates so that the steady-state probability distribution for each variable and the mean duration in each state of a variable would resemble their values in data that I have previously analyzed. I chose .1 and .05 for the rate (per year) of marital formation and marital dissolution, respectively. I

chose 5.0 and .4 for the annual rate of a man's entering and leaving employment, respectively, and 1.25 and .8 for the comparable annual rates for a woman. I set the monthly rates of friendship formation and dissolution for Y_A at .1 and .2, respectively, and at .2 and .4, respectively, for Y_B.

In examining interdependencies, I assumed that the above rates hold when the other variable equals unity (i.e., the person is married, employed or choses the other person as a friend). I assumed that the other rates were the following multiples of these:

$$\begin{aligned} r_{01} &= k_0 r_{23} \\ r_{02} &= k_1 r_{13} \\ r_{10} &= k_2 r_{32} \\ r_{20} &= k_3 r_{31} \end{aligned} \tag{33}$$

where

$$k = k_0 = k_1 = 1/k_3 = 1/k_4 \tag{34}$$

The two variables are independent when k equals unity. For each of the four pairs I let k = 1/4, 1/2, 1/1.25, 1, 1.25, 2, and 4. This gave a total of twenty-eight different bivariate processes.

V. FINDINGS

For each of the twenty-eight bivariate processes, there are two univariate processes that may be studied, giving a total of fifty-six cases. Moreover, for each case I considered several research designs, and for each

design I considered the true expected values and the biases for four kinds of outcomes: transition rates, the steady-state probability that the studied variable (Y_1) equals unity, the mean duration that this variable equals zero and unity, and the mean number of changes in a period (0,t). In all, these investigations produced far too many numbers to include in this chapter. So I summarize what I found and report only a small fraction of all numerical results for illustrative purposes.

First of all, the results indicate that the bias in each outcome is zero for every research design considered when k is unity, i.e., changes in Y_1 and Y_2 are independent. Because I considered only asymptotic results (i.e., results for an infinitely large sample size N) for maximum likelihood estimators, this must be true from a mathematical point of view. However, as a practical statement it provides evidence of the accuracy of the procedure used to calculate the expected value of each outcome and the bias obtained under different research designs.

Second, the calculations show that the more k differs from unity, the greater the percent bias in an outcome (defined as 100 times the bias divided by the true value). We might hypothesize that bias would increase absolutely as k differs more and more from unity, i.e., as the two variables become more interdependent. But this hypothesis ignores the fact that the expected <u>true</u> value of an outcome varies with k. For example, as k becomes larger, the mean duration that the studied variable Y_1 is in a given state may become smaller. In this situation, the absolute bias in the mean duration may decrease as k increases, even though the percent bias does increase, as we would expect. In most cases investigated, however, both absolute and percent biases increase as k increases.

Third, percent biases in predictions based on the analysis of panel data depend on the length of time between the waves of the panel, i.e., t. I arbitrarily chose two values for t, two and five ("years" for Pairs 1, 2 and 3; "months" for Pair 4), because these values approximate the interval in panel studies of these topics. The percent bias in the mean duration and in the mean number of changes in Y_1 in a period is often much larger when t is five rather than two. The opposite applies to the percent bias in predicting equilibrium probabilities; however, the interval t between waves of the panel affects percent bias for this outcome much less than the percent bias in the other outcomes. So I report illustrative results for t = 2 so that the panel design can be examined in its generally more favorable condition.

Fourth, results for the univariate model of Y_A, the more slowly changing variable, depend comparatively little on the type of data or the kind of sample. More importantly--and quite surprisingly--the findings show that both the absolute and percent biases in every outcome considered are usually small in magnitude when we study Y_A, the more slowly changing variable. When we study Y_B, the more rapidly changing variable, absolute and percent biases tend to be much larger than those for Y_A. An exception occurs in the case of Pair 4, in which the difference in the number of changes in Y_A and Y_B for a given interval (0,T) is comparatively small. As we might expect from this similarity, for Pair 4 percent bias is fairly similar whether we study Y_A and ignore Y_B or study Y_B and ignore Y_A. What is not intuitively obvious is that the percent biases for both Y_A and Y_B of Pair 4 tend to be relatively small. Thus, these investigations suggest that the percent biases resulting

from ignoring dynamic interdependencies are large mainly when the ignored variables change much more slowly than the variables studied.

Fifth, the pattern of biases associated with one of the four outcomes considered does not always give a good clue about the pattern of biases for another outcome, especially in studies of Y_B, the more rapidly changing variable. This necessitates a separate discussion of the biases associated with each of the four outcomes.

Transition Rates

It is somewhat difficult to think about a transition rate because we cannot observe it directly, any more than we can directly observe a probability of a change. Nevertheless, transition rates are conceptually important because they are the fundamental parameters of the stochastic process.

Table 1 reports the true rates of change in Y_1 when $Y_2 = 0$ (labelled "True, 0") and when $Y_2 = 1$ (labelled "True, 1") for selected cases. It also gives the estimated rates, r^*_{01} and r^*_{10}, of the univariate process for data from a two-wave panel with t = 2 ("P") and for event-history data ("E").[13] For each type of data I consider results for three kinds of samples: one in which $Y_2(0) = 0$ (labelled "0"), one in which $Y_2(0) = 1$ (labelled "1"), and a random sample from a population in equilibrium (labelled "r"). In Table 1 and in subsequent tables, the research design is denoted by the combination of these labels.

(Table 1 about here)

Several patterns are apparent in Table 1. Findings for the other cases not included in Table 1 are comparable, except as noted.

First, estimates of the rates for $Y_1 = Y_A$ depend comparatively little on the type of data or sample, as mentioned above. In addition, the estimates for all research designs fall between the two comparable true rates, as might be anticipated. They also tend toward the true rates where the omitted variable has the value that predominates in equilibrium (compare Table 2 below). For example, in Pair 3 the equilibrium probability that $Y_B = 1$ for a given value of Y_A is above 0.5, so the estimated rates for Y_A tend to be nearer those labelled "True, 1" than to those labelled "True, 0." This indicates that the relative rates of change in the omitted variable (Y_2) affect the estimated rates in the misspecified univariate model, as we might have anticipated.

Second, estimates of the rates for $Y_1 = Y_B$ vary considerably with the research design. Estimates for event-history designs are between the two comparable true rates, and the estimate for the random design ("Er") always lies between those for the two types of stratified designs ("E0" and "E1"), as might be expected. Furthermore, as in the case of $Y_1 = Y_A$, the estimated rates tend toward the true rates associated with the most prevalent value of the ignored variable in the steady-state.

On the other hand, estimated rates for a panel design are almost always smaller than those from the event-history design with the same kind of sample. Perhaps surprisingly, these estimates are sometimes smaller than either of the two comparable true rates (see Y_B of Pair 2, for example). And, unexpectedly, the estimate based on the random panel design ("Pr") does not always fall between the estimates for the two stratified panel designs (see Y_B of Pair 1, for example). It is not obvious why this occurs.

We might be tempted to conclude from these findings that estimates from an event-history design are better than those from a panel design when we are studying a variable that changes more rapidly than the variable being ignored (i.e., we are studying Y_B and ignoring Y_A). As we see below, this is usually the case, but not always, if "better" means smaller biases in the prediction of other outcomes.

Equilibrium Probabilities

Table 2 gives illustrative values of the equilibrium probability that $Y_1 = 1$ and the percent bias in predicting this probability using various research designs. The entries refer to the same pairs reported in Table 1. The bias for the misspecified model is obtained by inserting the estimates given in Table 1 in equation (9) and then subtracting the true value, which is calculated from (10) and the true rates.

(Table 2 about here)

Considering the comparatively poor estimates of rates obtained with the random panel design, it is somewhat surprising to find that the bias in the equilibrium probability that $Y_1 = 1$ is identically zero for this design. The lack of bias for this outcome and design apparently results from a mathematical identity rather than the coincidental choice of the transition rates of the bivariate process.

For both event-history and panel designs, biases in equilibrium probabilities are almost always greater for samples stratified by the omitted endogenous variable (Y_2) than for a random sample. Although we are accustomed to thinking that control of covariates improves the accuracy of estimates, the greater bias for predictions based on a

stratified sample is not so surprising upon reflection. The value of Y_2 upon which a sample is stratified is not a fixed characteristic of members of the sample. As t increases, the distribution of $Y_2(t)$ in a stratified sample approaches the distribution of $Y_2(0)$ in a random sample.[14]

It is reasonable to think that the predictions based on a random sample should resemble some kind of weighted average of the predictions in the two stratified samples (type 0 and type 1 samples). This suggests, in fact, that a prediction that weights the prediction from each type of stratified sample by the corresponding proportion in the steady-state might even have biases smaller than those in a random sample. To explore this possibility, Table 2 and subsequent tables also report percent biases for such weighted predictions (labelled "w").

In Table 2 we see that biases in the weighted predictions ("Pw" and "Ew") compare favorably in many cases to those from the random designs ("Pr" and "Er"). In fact, for Y_B, the more rapidly changing variable, there are very small biases in the equilibrium probability for the weighted panel prediction ("Pw"). This pattern also holds for Y_A when the interval between the waves of the panel is five years; however, when this interval is two years, the bias for the "Pw" design is sometimes worse than for either the "Ew" or "Er" designs. However, on the whole the percent bias in the equilibrium probability for the weighted prediction is almost as good as that for the random panel prediction and usually considerably better than either the random or weighted event-history prediction. Still, the percent biases for the latter predictions are

fairly small in most instances (especially in studies of Y_A). Thus panel analysis has a relative advantage over event-history analysis in predicting an equilibrium probability, but the size of the advantage is usually small.

Mean Durations

Table 3 reports the true mean duration that Y_1 equals zero and unity, assuming the process is in equilibrium, for the same pairs reported in the previous tables. It also reports the percent bias. Bias is obtained by inserting the estimates in Table 1 in (11), and then subtracting the true value, which is calculated from (12) and the true rates.

First note that the percent bias for the random event-history design is identically zero. In this design we are "predicting" the same quantity used to estimate the rates of the univariate model, so the bias must be zero.

(Table 3 about here)

Though the percent bias is occasionally smaller for one of the stratified samples than for the comparable weighted prediction, the difference in these instances tends to be comparatively small. On the other hand, biases are often considerably larger for the stratified samples than for the weighted prediction. This suggests one should ordinarily not predict mean duration from a stratified sample except as a step in calculating a weighted prediction.

The same reasoning leads to the conclusion that a weighted event-history prediction is preferable to either the random or weighted panel prediction. Occasionally the latter produce somewhat smaller biases than the former, but more often they are much worse. It is interesting that the weighted event-history prediction invariably has a negative bias, while the random panel prediction consistently has a positive bias. Thus the two predictions always bracket the true value.

Mean Number of Changes in a Period

Table 4 gives the expected value of the mean number of changes in Y_1 in a period, conditional on the initial value of Y_1.[15] It also reports the percent biases in these quantities for various research designs. Biases are obtained by inserting the estimates in Table 1 in equations (13) and (14), and then subtracting the true value, which is calculated from the true rates and equations (A1) and (A2) in the appendix.

(Table 4 about here)

As previously mentioned, for Y_A, the more slowly changing variable, both absolute and percent biases are fairly small for all designs. The random and weighted event-history predictions usually produce the smallest biases, but the random and weighted panel predictions are almost as good.

On the other hand, the biases for Y_B, the more rapidly changing variable, depend greatly on the research design. Biases for a panel design are almost always larger than those for the corresponding event-history design. The bias for the random panel design is usually the

largest in magnitude; furthermore, the bias for the weighted panel prediction cannot be relied on to be smaller than the bias for a prediction from a stratified panel design.

The magnitude of the bias is almost always smaller for the random event-history design than for any other design; often it is much smaller. Occasionally the bias for the weighted event-history prediction is the smallest, but in these instances it is only slightly smaller in size than the bias for the random event-history prediction. It should be noted that the random event-history prediction always has a negative bias while the weighted event-history prediction usually has a positive bias. Together these two predictions usually give a narrow bracket around the true expected value.

CONCLUSIONS

We cannot make general conclusions concerning the consequences of ignoring dynamic interdependencies among qualitative variables on the basis of the results for the small number of cases examined in this chapter. Yet, we can make a few general statements that can be regarded as hypotheses to be investigated more thoroughly in subsequent research.

As we would expect, biases from ignoring dynamic interrelationships tend to increase with the degree of interdependence of the variables. However, under some circumstances both absolute and percent biases are not large when interdependencies are ignored.

First of all, if the variable being studied changes more slowly than the ignored variable, biases are relatively small for all outcomes and research designs examined in this chapter. This is true even when the rates of changes in the variables differ by a factor as small

as two. Since changes in marital status are ordinarily much less frequent than changes in employment status, this suggests that research on marital stability that ignores changes in employment status can still reach sound results, but that research on employment stability that ignores changes in marital status is likely to err substantially, if dynamic interdependencies exist.

Secondly, even when other outcomes are not estimated very well, the equilibrium probability that a variable has a given value is estimated remarkably well, using either panel or event-history data to estimate a misspecified model. The panel prediction based on a random sample of a population in equilibrium equals the true value, in spite of the model misspecification. The prediction based on event-history analysis of a random sample is less accurate than the panel one, but is still quite good.

Third, biases in predicting mean durations and the number of changes in a variable in a period are often fairly small when based on estimates from event-history analysis of a random sample, even though the model estimated is misspecified. Often the bias in predictions of these quantities is under twenty percent, even when the comparable rates for the two values of the ignored variable differ by a factor of four. When a model is correctly specified, event-history analysis is preferable to panel analysis of qualitative variables for a variety of reasons (see, e.g., Tuma et al., 1979). This advantage apparently transfers to the case in which the model estimated ignores dynamic interdependencies among variables.

It is also interesting that for most outcomes event-history analysis of a sample distributed randomly across values of the ignored variable

tends to give consistently smaller biases (and fewer very large biases) than event-history analysis of samples stratified on values of the omitted endogenous variable. Since the latter procedure attempts to control partially for the effects of the omitted variable, this may seem somewhat surprising at first. However, it seems reasonable when we realize that the omitted variables are not fixed characteristics but change over time until eventually they reach the distribution in a random sample.

In sum, the findings of the research reported in this chapter suggest some guidelines for studying change in dynamically interdependent qualitative variables, even when we estimate a model that ignores interdependence. But definitive conclusions are not yet possible, and further research is acutely needed to extend and explain the findings of this chapter.

APPENDIX

For the bivariate time-homogeneous Markov process, the number of changes to state s before t when $Y_1(0) = y_1$, $n_{y_1 s}(t)$, is given by:

$$n_{y_1 y_1}(t) = \sum_{j=0}^{1} \sum_{k=0}^{1} \gamma_{y_1(1-j)}\, \gamma_{(1-y_1)(1-k)} \left[\frac{\mu_{y_1 j} \mu_{(1-y_1)k}\, t}{Q(0)} + \sum_{i=1}^{3} \frac{(\mu_{y_1 j} + \zeta_i)(\mu_{(1-y_1)k} + \zeta_i)(e^{\zeta_i t} - 1)}{\zeta_i Q(\zeta_i)} \right] \tag{A1}$$

$$n_{y_1(1-y_1)}(t) = \sum_{j=0}^{1} \gamma_{y_1(1-j)} \left[\frac{\mu_{y_1 j} \mu_{j y_1} \mu_{(1-y_1)(1-y_1)}\, t}{Q(0)} + \sum_{i=1}^{3} \frac{(\mu_{y_1 j} + \zeta_i)(\mu_{j y_1} + \zeta_i)(\mu_{(1-y_1)(1-y_1)} + \zeta_i)(e^{\zeta_i t} - 1)}{\zeta_i Q(\zeta_i)} \right] \tag{A2}$$

where for $k = 0, 1$:

$$\gamma_{jk} = (k - \beta_j)\mu_{jk} \tag{A3}$$

$$\mu_{y_1 k} = \tfrac{1}{2}\left[-(r_a + r_b) + (-1)^k \sqrt{(r_a - r_b)^2 + 4 r_{ab} r_{ba}}\,\right] \tag{A4}$$

$$\beta_{y_1} = \left[r_{ab} + r_{ba} + p_{y_1}(\infty) r_{bd} + (1 - p_{y_1}(\infty)) r_{ac} + \mu_{y_1 0}\right] / (\mu_{y_1 0} - \mu_{y_1 1}) \tag{A5}$$

$$a = 2y_1 + 1 \tag{A6}$$

$$b = 2y_1 \tag{A7}$$

$$c = 2(1-y_1) + 1 \tag{A8}$$

$$d = 2(1-y_1) \tag{A9}$$

$$Q(\zeta) = \zeta(\zeta-\zeta_1)(2\zeta-\zeta_2-\zeta_3) + (\zeta-\zeta_2)(\zeta-\zeta_3)(2\zeta-\zeta_1) \tag{A10}$$

and ζ_i, $i = 1, 2, 3$, are the distinct real roots[16] of

$$\zeta^3 + \xi_2\zeta^2 + \xi_1\zeta + \xi_0 = 0 \tag{A11}$$

where

$$\xi_2 = \mu_{00} + \mu_{01} + \mu_{10} + \mu_{11} \tag{A12}$$

$$\xi_1 = \mu_{00}\mu_{01} + \mu_{11}\mu_{10} + (\mu_{00} + \mu_{01})(\mu_{10} + \mu_{11}) - (\gamma_{00} + \gamma_{01})(\gamma_{10} + \gamma_{11}) \tag{A13}$$

$$\xi_0 = \mu_{00}\mu_{01}(\mu_{10} + \mu_{11} - \gamma_{10} - \gamma_{11}) + \mu_{11}\mu_{10}(\mu_{01} + \mu_{00} - \gamma_{01} - \gamma_{00}) \tag{A14}$$

FOOTNOTES

[1]If we know the form or structure of the true model, we may also choose to use our incomplete data to make inferences about parameters in the true model. The literature on this problem is scanty, especially for qualitative variables. Though research on this topic is needed, I do not pursue it here.

[2]Simon and Ando (1961) and Ando and Fisher (1963) (summarized in Fisher and Ando (1962)) have considered the problem for a dynamic system of interdependent quantitative variables, in particular, for a system of linear differential equations. When the system is in equilibrium, this model reduces to a system of simultaneous linear equations. The consequences of ignoring an equation in this system are well-known; see, for example, Johnston, 1972.

[3]In this paper the symbol "$\triangleq$" means "equals by definition."

[4]For brevity I sometimes refer to transition rates (Coleman, 1964) simply as rates; other sociologists have called them transition intensities (e.g., Holland and Leinhardt, 1977), infinitesimal generators (e.g., Mayer, 1972), and "q's" (e.g., Coleman, 1964; Singer and Spilerman, 1974, 1976).

[5]The additional assumptions are that probabilities range between zero and unity:

$$0 \leq p_{sy}(u,t) \leq 1$$

for all u and t and for all y and s, and that there are no discontinuities in transition probabilities:

$$\lim_{\Delta t \to 0} P(t, t + \Delta t) = I$$

where I is the identity matrix, and that the system is closed:

$$\sum_{y} p_{sy}(u,t) = 1$$

for all s.

[6]Several ways of relaxing these assumptions are available. The population homogeneity assumption can be relaxed by assuming that transition rates have some specified probability distribution in the population (e.g., Blumen et al., 1955; Silock, 1954; Spilerman, 1972a; Tuma, 1978), or that transition rates depend on exogenous observed characteristics of members of the population in some specified way (e.g., Coleman, 1964; 1973; Spilerman, 1972b; Tuma, 1976; Hannan et al., 1977; Tuma et al., 1979). The time-independence assumption can be relaxed by assuming that transition rates have some specified functional dependence on time (e.g., Mayer, 1972; Sørensen, 1975; Tuma, 1976), or that transition rates vary freely from one specified time period to another (e.g., Tuma et al., 1979).

[7]For a finite-state, time-stationary Markov process, a steady-state probability distribution exists if states intercommunicate, i.e., it is possible to move from each state to every other state in a finite number of moves (Cox and Miller, 1965). Both the bivariate and univariate models meet this condition.

[8]One eigenvalue is necessarily zero; this is defined to be λ_1. For many sociological applications, the other eigenvalues are distinct, negative, and nonzero, as I assume here. However, this assumption is not always true; in these instances (10) has a different form, which I do not consider.

[9] In a system of M dynamically interdependent dichotomous variables, cross-sectional data provide information on $Y_m(t)$, m = 1,.., M, for the N members of the sample and thus allow 2^M-1 linearly independent proportions to be calculated. However, $2^M M$ transition rates describe the change process. Because $2^M M > 2^M-1$ for M≥1, cross-sectional data do not provide sufficient information for the transition rates to be estimated uniquely, even if the process is in equilibrium and the model is specified correctly. Consequently, I do not discuss cross-sectional data.

[10] When (22) does not hold, the data are not consistent with any time-homogeneous, continuous-time, univariate Markov process.

[11] This solution of equation (7) holds when λ_j, j = 2, . . , M, are distinct, negative eigenvalues of the matrix of transition rates (Bartholomew, 1973).

[12] If we considered sample size as a varying feature of the research design, we might be most interested in the mean squared errors associated with the misspecified model. The mean squared error depends on variance as well as bias. Because I am concentrating on situations in which the sample size is large, I ignore variance. In effect, I assume variance is zero.

[13] See equations (23), (24), (31), and (32).

[14] This suggests that inferences from stratified and random samples differ most when the interval between waves of a panel is comparatively short.

[15] Recall that the derivation of the true expected number requires the assumption that the population is in equilibrium.

[16] Let $\phi = \xi_1/3 - \xi_2^2/9$ and $\psi = \xi_1\xi_2/6 - \xi_0/2 - \xi_2^3/27$. The roots of (A11) are real and distinct if $\phi^3 + \psi^2 < 0$. In this case, the roots are

$$\zeta_1 = [2\sqrt{-\phi}\ \cos(\theta/3)] - \xi_2/3$$

$$\zeta_2 = [2\sqrt{-\phi}\ \cos(\theta/3 + 120^\circ)] - \xi_2/3$$

$$\zeta_3 = [2\sqrt{-\phi}\ \cos(\theta/3 + 240^\circ)] - \xi_2/3$$

where

$$\theta = \cos^{-1}(\Psi/\sqrt{-\phi^3})$$

In the applications that I have considered, the roots of (A11) have always been real, distinct, and negative. However, it is not clear what restrictions on the transition rates are necessary for this to occur.

REFERENCES

Albert, A.

1962 "Estimating the infinitesimal generator of a continuous time, finite state Markov process." Annals of Mathematical Statistics 33 (June): 727-753.

Ando, A. and Fisher, F. M.

1963 "Near-decomposability, partition and aggregation, and the relevance of stability discussions." International Economic Review (January): 53-67.

Bartholomew, David J.

1973 Stochastic Models for Social Processes. Second edition. New York: Wiley.

Blumen, Isadore, Kogan, M. and McCarthy, Philip J.

1955 The Industrial Mobility of Labor as a Probability Process. (Cornell Studies in Industrial and Labor Relations, No. 6.) Ithaca, NY: Cornell University Press.

Coleman, James S.

1964 Introduction to Mathematical Sociology. New York: Free Press.

1973 The Mathematics of Collective Action. Chicago: Aldine Publishing.

Cox, D. R. and Miller, H. D.

1965 The Theory of Stochastic Processes. London: Methuen and Company.

Feller, William

1968 An Introduction to Probability Theory and Its Application. Volume I, third edition. New York: Wiley.

Fisher, F. M. and Ando, A.

1962 "Two theorems of ceteris paribus in the analysis of dynamic systems." American Political Science Review, Vol. 56: 108-113.

Hannan, Michael T., Tuma, Nancy Brandon and Groeneveld, Lyle

1977 "Income and marital events: evidence from an income-maintenance experiment." American Journal of Sociology 82 (May): 1186-1211.

Holland, Paul W. and Leinhardt, Samuel

1977 "A dynamic model for social networks." Journal of Mathematical Sociology, Vol. 5: 5-20.

Johnston, J.

1972 Econometric Methods. New York: McGraw-Hill.

Mayer, Thomas F.

1972 "Models of intragenerational mobility." Pp. 308-357 in J. Berger, M. Zelditch, Jr., and B. Anderson (eds.), Sociological Theories in Progress. Vol. 2. Boston: Houghton Mifflin Company.

Silock, H.

1954 "The phenomenon of labour turnover." Journal of the Royal Statistical Society A117: 429-440.

Simon, H. A. and Ando, A.

1961 "Aggregation of variables in dynamic systems." Econometrica 29 (April): 111-138.

Singer, Burton and Spilerman, Seymour

1974 "Social mobility models for heterogeneous populations." In Herbert L. Costner (ed.), Sociological Methodology 1973-1974. San Francisco: Jossey-Bass.

1976 "The representation of social processes by Markov models." American Journal of Sociology 82 (July): 1-54.

Sørensen, Aage B.

1975 "The structure of intragenerational mobility." American Sociological Review 40: 456-471.

Sørensen, Aage B. and Hallinan, Maureen T.

1975 "A stochastic model for change in group structure." Social Science Research 5: 43-61.

Spilerman, Seymour

1972a "Extensions of the mover-stayer model." American Journal of Sociology 78: 599-627.

1972b "The analysis of mobility processes by the introduction of independent variables into a Markov chain." American Sociological Review 37 (June): 277-294.

Tuma, Nancy Brandon

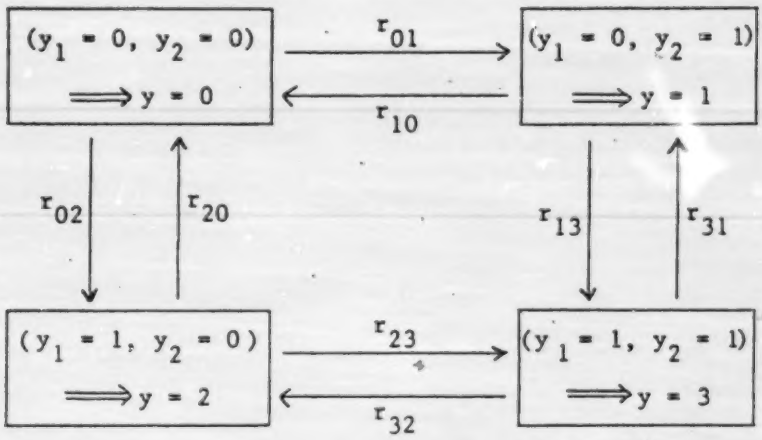
1976 "Rewards, resources, and the rate of mobility: a nonstationary multivariate stochastic model." American Sociological Review 41: 338-360.

1978 "Effects of labor market structure on job-shift patterns." Paper presented at the annual American Sociological Association meetings, San Francisco, Calif.

Tuma, Nancy Brandon, Hannan, Michael T. and Groeneveld, Lyle

1979 "Dynamic analysis of event histories." American Journal of Sociology 84 (January). In press.

(a) Bivariate Markov Model of $Y(t) = \{Y_1(t), Y_2(t)\}$



(b) Univariate Markov Model of $Y_1(t)$

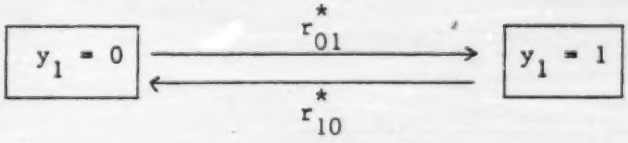


Figure 1a. Bivariate and univariate Markov models of change in a dichotomous variable Y_1 that is dynamically related to another dichotomous variable Y_2.

TABLE 1

Transition Rates: True Values for a Bivariate Model and Expected Values for a Misspecified Univariate Model of Y_1 Estimated by Various Research Designs

	Y_1=Y_A		Y_1=Y_B	
	Change from y_1=0 to y_1=1	Change from y_1=1 to y_1=0	Change from y_1=0 to y_1=1	Change from y_1=1 to y_1=0
Pair 1				
True, 0	.050	.100	2.500	.800
P0	.082	.057	2.230	.628
Pr	.088	.054	1.324	.212
P1	.090	.053	3.020	.291
E0	.087	.054	2.524	.727
Er	.088	.054	3.012	.490
E1	.088	.054	4.906	.421
True, 1	.100	.050	5.000	.400
Pair 2				
True, 0	.200	.025	2.500	.400
P0	.127	.037	1.848	.400
Pr	.113	.040	.947	.455
P1	.111	.042	1.245	.750
E0	.117	.040	2.330	.[illegible]
Er	.113	.040	1.329	.615
E1	.113	.040	1.262	.764
True, 1	.100	.050	1.250	.800
Pair 3				
True, 0	2.500	.400	10.000	.200
P0	1.280	.767	4.022	.224
Pr	1.267	.766	3.568	.202
P1	1.268	.767	4.080	.233
E0	1.408	.723	8.418	.302
Er	1.272	.768	5.446	.318
E1	1.270	.772	5.155	.329
True, 1	1.250	.800	5.000	.400
Pair 4				
True, 0	.050	.400	.100	.800
P0	.053	.368	.105	.755
Pr	.054	.325	.104	.621
P1	.073	.233	.160	.458
E0	.055	.350	.108	.743
Er	.053	.323	.110	.624
E1	.058	.280	.127	.473
True, 1	.100	.200	.200	.400

TABLE 2

The Equilibrium Probability that Y_1=1: True Value for a Bivariate Model and Percent Bias for a Misspecified Univariate Model of Y_1 Estimated by Various Research Designs

	Pair 1		Pair 2		Pair 3		Pair 4	
	Y_1=Y_A	Y_1=Y_B	Y_1=Y_A	Y_1=Y_B	Y_1=Y_A	Y_1=Y_B	Y_1=Y_A	Y_1=Y_B
True	.621	.862	.739	.676	.623	.946	.143	.143
P0	-5.0	-9.5	4.9	21.6	.4	.1	-11.4	-14.7
P1	.9	5.8	-1.9	-7.6	-.0	-.1	56.3	81.3
Pw	.1	.0	.3	.0	.0	-.0	-1.8	-1.0
Pr	0	0	0	0	0	0	0	0
E0	-.8	-10.0	.9	24.4	6.0	3.0	-4.9	-11.4
E1	.1	6.8	-.3	-7.8	-.2	-.5	19.9	48.5
Ew	-.0	.5	.1	.6	.1	.8	-1.4	-2.8
Er	-.0	-.2	-.0	1.2	.0	-.1	2.3	4.6

TABLE 3

Mean Durations: True Values for a Bivariate Model and Percent Biases for a Misspecified Univariate Model of Y_1 Estimated by Various Research Designs

	$Y_1=Y_A$		$Y_1=Y_B$	
	$Y_1=0$	$Y_1=1$	$Y_1=0$	$Y_1=1$
Pair 1				
True	11.397	18.632	.332	2.042
P0	6.4	-6.4	35.1	-21.9
P1	-1.9	-.6	-.3	68.1
Pw	-.8	-.4	13.1	34.0
Pr	.0	.1	127.4	131.2
E0	1.1	-.8	19.3	-32.6
E1	-.4	.1	-38.[illegible]	16.3
Ew	-.2	-.1	-16.6	-2.2
Er	0	0	0	0
Pair 2				
True	8.818	24.893	.752	1.627
P0	-11.0	8.3	28.0	33.6
P1	1.9	-5.0	6.8	-18.0
Pw	-2.2	-.6	-2.3	.7
Pr	.1	.3	40.3	35.2
E0	-2.8	.7	-43.0	39.0
E1	.4	-.5	5.4	-19.5
Ew	-.6	-.1	-7.3	-4.2
Er	0	0	0	0
Pair 3				
True	.786	1.301	.184	3.146
P0	-.6	.2	35.4	42.1
P1	.3	.2	33.5	36.3
Pw	.3	.2	34.2	38.5
Pr	.4	.4	52.6	57.4
E0	-9.7	6.4	-35.3	5.2
E1	.2	-.5	5.6	-3.4
Ew	-.3	-.1	-9.8	-.1
Er	0	0	0	0
Pair 4				
True	18.081	3.095	9.127	1.603
P0	3.6	-12.3	4.6	-17.4
P1	-24.1	27.5	-31.6	36.2
Pw	-.3	-6.6	-.6	-9.7
Pr	2.0	-.7	5.9	.5
E0	.6	-7.7	1.7	-16.0
E1	-4.5	15.4	-13.9	32.0
Ew	-.2	-4.4	-.5	-9.0
Er	0	0	0	0

TABLE 4

Mean Number of Changes in Y_1 in (0,2): True Values for a Bivariate Model and Percent Biases for a Misspecified Univariate Model of Y_1 Estimated by Various Research Designs*

	Y_1-Y_A			Y_1-Y_B		
	$n_{y_1y_1}(2)$	$n_{01}(2)$	$n_{10}(2)$	$n_{y_1y_1}(2)$	$n_{01}(2)$	$n_{10}(2)$
Pair 1						
True	.009	.162	.102	.754	1.609	.896
P0	.5	-5.6	6.1	7.3	-1.4	12.8
P1	1.3	1.8	-.5	-40.1	-15.3	-66.2
Pw	1.2	.8	.4	-22.2	-10.0	-23.7
Pr	-.2	-.1	-.1	-66.6	-33.2	-133.6
E0	-.3	-1.0	.8	26.8	7.6	24.0
E1	.2	.3	-.1	-6.8	.9	-14.6
Ew	.2	.1	.1	6.0	3.5	3.9
Er	-.0	-.0	-.0	-4.2	-1.7	-3.9
Pair 2						
True	.008	.204	.078	.656	1.317	.971
P0	2.7	10.8	-8.2	-21.8	.7	-40.8
P1	3.1	-1.7	4.7	7.6	.1	9.7
Pw	2.9	2.3	.9	-.1	.2	.3
Pr	-.5	-.1	-.3	-37.7	-20.8	-36.1
E0	1.8	2.5	-.7	-7.0	9.9	-26.3
E1	-.1	-.4	.4	9.8	1.2	11.0
Ew	.5	.5	.0	5.4	3.4	3.6
Er	-.1	-.0	-.0	-4.1	-1.4	-3.4
Pair 3						
True	.728	1.340	1.098	.550	1.495	.605
P0	.2	.3	-.0	-32.0	-11.6	-41.9
P1	-.3	-.2	-.2	-29.1	-10.6	-36.3
Pw	-.3	-.2	-.2	-30.2	-11.0	-38.3
Pr	-.6	-.3	-.4	-39.7	-14.6	-57.2
E0	.9	3.4	-2.7	-.1	1.3	-3.6
E1	.2	.0	.3	2.1	.5	2.7
Ew	.3	.2	.2	1.3	.8	.4
Er	-.0	-.0	-.0	-.3	-.1	-.3
Pair 4						
True	.029	.106	.486	.092	.205	.752
P0	5.5	-3.4	8.0	5.0	-4.4	8.0
P1	4.7	29.3	-20.8	10.3	38.8	-19.9
Pw	5.4	1.2	4.8	5.7	1.8	4.8
Pr	-3.0	-2.1	-.4	-8.6	-6.0	-1.6
E0	4.3	-.7	4.7	6.6	-2.0	7.4
E1	-8.6	4.1	-12.1	-8.6	13.0	-18.5
Ew	2.4	.0	2.6	4.4	.2	4.4
Er	-1.7	-.2	-.9	-3.3	-.8	-1.2

*See text for definitions of $n_{y_1y_1}(z)$, $n_{01}(z)$, and $n_{10}(z)$.